

Funciones de Muestreo en Python

Se le conoce como muestreo a la técnica para la selección de una muestra a partir de una población estadística.

Al elegir una muestra aleatoria se espera conseguir que sus propiedades sean extrapolables a la población. Este proceso permite ahorrar recursos, y a la vez obtener resultados parecidos a los que se alcanzarían si se realizase un estudio a toda la población. En las investigaciones llevadas por empresarios y de la medicina se usa muestreo extensivamente en recoger información sobre poblaciones.

Fuente:

[https://es.wikipedia.org/wiki/Muestreo_\(estad%C3%ADstica\)](https://es.wikipedia.org/wiki/Muestreo_(estad%C3%ADstica))

Tabla de datos de economía y turismo en el Centro Histórico CDMX

```
import os

def import_or_install(package):
    try:
        __import__(package)
    except ImportError:
        os.system("pip install " + package)
```

```
import_or_install('pandas')
import_or_install('numpy')
import_or_install('random')
import_or_install('wget')
import_or_install('io')
```

```
import pandas as pd
import numpy as np
import random
import wget
import io
```

Leemos la base de datos a trabajar en nuestra ruta de origen

```
!wget -O datos.csv https://datos.cdmx.gob.mx/dataset/d19d49ea-8a73-4bf5-910e-81060068bd3f/resource/dl
econdata = pd.read_csv("datos.csv")
```

```
--2023-09-27 05:25:11-- https://datos.cdmx.gob.mx/dataset/d19d49ea-8a73-4bf5-910e-81060068bd3f/resource/dl
Resolving datos.cdmx.gob.mx (datos.cdmx.gob.mx)... 45.60.240.232
Connecting to datos.cdmx.gob.mx (datos.cdmx.gob.mx)|45.60.240.232|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 57045 (56K) [application/octet-stream]
Saving to: 'datos.csv'
```

```
datos.csv          100%[=====>]  55.71K  --.-KB/s    in 0.1s
```

econdata.head()

	id int64	geo_point_2d object	geo_shape object	clave_cat object	delegacion object	perimetro object
0	0	19.424781053,-9...	{"type": "Polygon"...	307_130_11	Cuauhtémoc	B
1	1	19.4346139576,-...	{"type": "MultiPoi...	002_008_01	Cuautémoc	A
2	2	19.4340695945,-...	{"type": "MultiPoi...	006_002_12	Cuautémoc	A
3	3	19.42489472,-99...	{"type": "MultiPoi...	323_102_06	Venustiano Carra...	B
4	4	19.42358238,-99...	{"type": "MultiPoi...	323_115_12	Venustiano Carra...	B

5 rows, showing 10 per page << < Page 1 of 1 > >>

Funciones de muestreo:

1) Muestreo aleatorio Simple:

Es considerado el método más sencillo. Mediante una tabla de números al azar se eligen las zonas que se quieren muestrear. Este tipo de muestreo posee algunos inconvenientes. Por un lado, supone definir de antemano los límites de un yacimiento, y no siempre se conocen con certeza. Por otro lado, el carácter aleatorio de las tablas numéricas provoca que en algunas áreas se acumulen las muestras, mientras que en otras permanecen intactas.

Fuente:

https://es.wikipedia.org/wiki/Estrategias_de_muestreo

aleat_8 = econdata.sample(n=8)
aleat_8

	id int64	geo_point_2d object	geo_shape object	clave_cat object	delegacion object	perimetro object
169	169	19.4365391003,-...	{"type": "MultiPoi...	004_093_16	Cuautémoc	A
148	148	19.4388349207,-...	{"type": "MultiPoi...	012_147_13	Cuautémoc	B
173	173	19.4314834886,-...	{"type": "MultiPoi...	006_026_38	Cuautémoc	A
224	224	19.4238397797,-...	{"type": "MultiPoi...	001_109_11	Cuautémoc	B
20	20	19.4357307042,-...	{"type": "MultiPoi...	004_098_26	Cuautémoc	A
13	13	19.43878785,-99...	{"type": "MultiPoi...	012_147_02	Cuautémoc	B
29	29	19.4291936373,-...	{"type": "MultiPoi...	323_145_10	Venustiano Carra...	B
34	34	19.438545231,-9...	{"type": "MultiPoi...	003_103_23	Cuautémoc	A

8 rows, showing 10 per page << < Page 1 of 1 > >>

```
aleat_8_2 = econdata.sample(n=8)
aleat_8_2
```

	id int64	geo_point_2d object	geo_shape object	clave_cat object	delegacion object	perimetro object
126	126	19.42774805,-99...	{"type": "MultiPoi...	006_078_04	Cuautémoc	A
123	123	19.4378770032,-...	{"type": "MultiPoi...	004_086_36	Cuautémoc	A
180	180	19.4357633849,-...	{"type": "MultiPoi...	004_094_32	Cuautémoc	A
22	22	19.4348361174,-...	{"type": "MultiPoi...	002_015_01	Cuautémoc	B
116	116	19.43339234,-99...	{"type": "MultiPoi...	002_017_02	Cuautémoc	B
215	215	19.4383767258,-...	{"type": "MultiPoi...	004_082_16	Cuautémoc	A
212	212	19.42155871,-99...	{"type": "MultiPoi...	001_095_08	Cuautémoc	B
67	67	19.4434657626,-...	{"type": "MultiPoi...	005_057_01	Cuautémoc	B

8 rows, showing 10 per page

<< < Page 1 of 1 > >>

Proporción al 25%

```
prop_25 = econdata.sample(frac = .25)
prop_25.head()
```

	id int64	geo_point_2d object	geo_shape object	clave_cat object	delegacion object	perimetro object
68	68	19.4273142523,-...	{"type": "MultiPoi...	006_082_02	Cuautémoc	A
22	22	19.4348361174,-...	{"type": "MultiPoi...	002_015_01	Cuautémoc	B
19	19	19.4317119617,-9...	{"type": "MultiPoi...	006_026_28	Cuautémoc	A
205	205	19.435760727,-9...	{"type": "MultiPoi...	005_129_16	Cuautémoc	A
139	139	19.4331109726,-...	{"type": "MultiPoi...	006_009_11	Cuautémoc	A

5 rows, showing 10 per page

<< < Page 1 of 1 > >>


2) Muestreo sistemático:

Muestreo sistemático Este tipo de muestreo se basa en el empleo de una red de lugares equidistantes, como por ejemplo eligiendo un cuadrado cada dos. Uno de los problemas que tiene es que es muy probable error, ya que este método de espaciado regular corre el riesgo de errar (o acertar) todas las muestras sin excepción si la distribución misma es también constante. **Fuente:**

https://es.wikipedia.org/wiki/Estrategias_de_muestreo

```
def systematic_sampling(econdata, step):
    indexes = np.arange(0, len(econdata), step=step)
    systematic_sample = econdata.iloc[indexes]
    return systematic_sample

systematic_sample = systematic_sampling(econdata, 3)
systematic_sample
```

	id int64 0 - 228 	geo_point_2d object 19.42478105... - 1.3% 75 others 97.4% Missing 1.3%	geo_shape object {"type": "Pol... - 1.3% 75 others 97.4% Missing 1.3%	clave_cat object 307_130_11 1.3% 323_102_06 1.3% 75 others 97.4%	delegacion object Cuautémoc 76.6% Venustiano 13% Cuauhtémoc 10.4%	perimetro object B 58.4 A 41.6
0	0	19.424781053,-9...	{"type": "Polygon"...	307_130_11	Cuauhtémoc	B
3	3	19.42489472,-99....	{"type": "MultiPoi...	323_102_06	Venustiano Carra...	B
6	6	19.43553422,-99...	{"type": "MultiPoi...	318_116_11	Venustiano Carra...	B
9	9	19.4407152937,-...	{"type": "MultiPoi...	012_146_22	Cuautémoc	B
12	12	19.43990186,-99....	{"type": "MultiPoi...	003_079_16	Cuautémoc	B
15	15	19.42413788,-99....	{"type": "MultiPoi...	307_153_11	Cuautémoc	B
18	18	19.4331161255,-9...	{"type": "MultiPoi...	006_021_01	Cuautémoc	A
21	21	19.43614459,-99....	{"type": "MultiPoi...	004_098_01	Cuautémoc	A
24	24	19.4285279152,-...	{"type": "MultiPoi...	002_067_19	Cuautémoc	B
27	27	19.4348360773,-...	{"type": "MultiPoi...	002_016_01	Cuautémoc	B

77 rows, showing 10 per page << < Page 1 of 8 > >>

3) Muestreo Estratificado:

Este tipo de muestreo se caracteriza por la combinación de elementos de los otros tipos de muestreo: muestreo aleatorio simple, aleatorio estratificado y sistemático. Es un intento de reducir la arbitrariedad en la toma de muestras.

Fuente:

https://es.wikipedia.org/wiki/Estrategias_de_muestreo

```
econdata['estratificado'] = econdata['delegacion'] + "," + econdata['tipo']
(econdata['estratificado'].value_counts()/len(econdata)).sort_values(ascending = False)
```

```
Cuautémoc,Hotel          0.643478
Cuautémoc,Museo          0.156522
Venustiano Carranza,Hotel 0.078261
Cuauhtémoc,Mercado       0.073913
Venustiano Carranza,Mercado 0.047826
Name: estratificado, dtype: float64
```

Nuestros datos dicen que la proporción es la siguiente:

- 1. Hoteles en Cuauhtémoc es de 0.5
- 2. Museos en Cuauhtémoc es de 0.2
- 3. Hoteles en Venuestiano Carranza es de 0.1
- 4. Mercados en Cuauhtémoc es de 0.1
- 5. Mercados en Venuestiano Carranza es de 0.1

```
def data_estratificada(econdata, nombres_columnas_estrat, valores_estrat, prop_estrat, random_state=None):
    df_estrat = pd.DataFrame(columns = econdata.columns) # Creamos un data frame vacío con los nombres de las columnas
    pos = -1
    for i in range(len(valores_estrat)): # iteración sobre los valores estratificados
        pos += 1
        if pos == len(valores_estrat) - 1:
            ratio_len = len(econdata) - len(df_estrat) # si es la iteración final calcula el número de filas restantes
        else:
            ratio_len = int(len(econdata) * prop_estrat[i]) # calcula el número de filas según la proporción
        df_filtrado = econdata[econdata[nombres_columnas_estrat] == valores_estrat[i]] # filtra los datos por el valor estratificado
        df_temp = df_filtrado.sample(replace=True, n=ratio_len, random_state=random_state) # haz un sampleo con reposición
        df_estrat = pd.concat([df_estrat, df_temp]) # junta las tablas de sample con la estratificación

    return df_estrat # Return the stratified re-sampled data
```

```
valores_estrat = ['Cuautémoc,Hotel', 'Cuautémoc,Museo', 'Venustiano Carranza,Hotel', 'Cuauhtémoc,Mercado', 'Cuauhtémoc,Parque']
prop_estrat = [0.5, 0.2, 0.1, 0.1,0.1]
df_estrat = data_estratificada(econdata, 'estratificado', valores_estrat, prop_estrat, random_state=42)
df_estrat
```

	id object	geo_point_2d object	geo_shape object	clave_cat object	delegacion object	perimetro object
	27 2.2%	19.4348360... 2.2%	{"type": "Mul... 2.2%	002_016_01 2.2%	Cuautémoc 70%	
	163 2.2%	138 others 96.5%	138 others 96.5%	002_059_01 2.2%	Venustiano 20%	B 61.7
	138 others 95.7%	Missing 1.3%	Missing 1.3%	134 others 95.7%	Cuauhtémoc 10%	A 38.3
164	164	19.4388741511,-9...	{"type": "MultiPoi...	003_113_03	Cuautémoc	B
142	142	19.4263681354,-...	{"type": "MultiPoi...	006_127_14	Cuautémoc	A
27	27	19.4348360773,-...	{"type": "MultiPoi...	002_016_01	Cuautémoc	B
168	168	19.4349726565,-...	{"type": "MultiPoi...	002_014_23	Cuautémoc	B
113	113	19.43374405,-99...	{"type": "MultiPoi...	001_012_13	Cuautémoc	A
34	34	19.438545231,-9...	{"type": "MultiPoi...	003_103_23	Cuautémoc	A
164	164	19.4388741511,-9...	{"type": "MultiPoi...	003_113_03	Cuautémoc	B
191	191	19.43985567,-99...	{"type": "MultiPoi...	003_079_12	Cuautémoc	B
117	117	19.4253041176,-...	{"type": "MultiPoi...	001_078_03	Cuautémoc	B
137	137	19.4421185698,-...	{"type": "MultiPoi...	012_108_03	Cuautémoc	B

230 rows, showing 10 per page

<< < Page 1 of 23 > >>

Referencias

<https://platzi.com/clases/3140-estadistica-inferencial-python/49505-funciones-de-muestreo-en-python/>

https://es.wikipedia.org/wiki/Estrategias_de_muestreo

[https://es.wikipedia.org/wiki/Muestreo_\(estad%C3%ADstica\)](https://es.wikipedia.org/wiki/Muestreo_(estad%C3%ADstica))

<https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet>